Three alternative architectures of digital ratioed compressor design with application to inner-product processing

C.-C.Wang, P.-M.Lee and C.-J.Huang

Abstract: Inner-product calculations are often required in digital neural computing. The critical path of the inner product of two binary vectors is the carry propagation delay generated from individual product terms. Three alternative architecture for arranging digital ratioed compressors are presented, to reduce the carry propagation delay in the critical path wherein an improved design of a 3-2 compressor is used to serve as the basic building element. The carry propagation delay estimation for the there architectures is also derived and compared. The theoretical analyses and Verilog simulations both indicate that one of the architectures presented might offer a sub-optimal solution for summing the individual product terms in the inner-product computation. Furthermore, a real chip for the sub-optimal architecture was fabricated and fully tested. The testing results prove the correctness of its functions and performance.

1 Introduction

Many attempts have been made at the realisation of neural networks, mainly owing to their attractive pattern recognition features, [1, 2]. In the computation of neural networks, the inner product of two vectors might be one of the most frequently used mathematical operations. Unavoidably, the carry propagation will occur if the neural networks are dedicated to either discrete or digital signals. For instance, the recall of pattern pairs stored in the discrete bidirectional associative memory (BAM) needs to compute a summation in the form of $Y=th(\sum_{i=1}^{n} Y_i \cdot (X_i \cdot X))$, where X is the input pattern, Y is the output pattern, X_is and Y_is are stored pattern pairs and $th(\cdot)$ is a threshold function. Notably, the components of every vector are either bipolar or binary. If n is large in the above calculation, then the carry propagation of the inner product of the vectors will likely cause a critical delay in the entire neural computation.

Since neural computing is composed of a massive amount of inner product calculations, the need to shorten the delay becomes urgent. Many high-speed logic design styles have been announced, however these logics suffer from different difficulties. For example, domino logic [3] cannot be non-inverting; NORA [4] has the charge-sharing problem; all-N-logic [5] and robust single-phase clocking [6] cannot operate correctly under clocks with a short rise time or fall time, which cannot be easily integrated with other parts of the logic design; and single-phase logic [7] and Zipper CMOS [8] contain slow P-logic blocks. Complementary pass-transistor logic (CPL) proposed by Yano *et al.* [9] is twice as fast as conventional CMOS, whereas it needs a larger area of silicon, like the conventional CMOS, due to the mixed interconnection. Moreover, the noise margin and speed degradation, caused by the mismatched input signal level and the logic threshold voltage of the CMOS driver, needs to be taken into consideration when CPL is implemented. Although Zhang *et al.* proposed a so-call C^2PL (complex CPL) and demonstrated that the problems of CPL are all saved in [10], several physical design factors are not fully considered or implemented. First, the sizes of the NMOS transistors for pass logics cannot be minimal secondly the driving inverters' sizes have to be properly tuned, thirdly, the original design of [10] not only gives a poor fan-in and fan-out capability, but also produces very asymmetrical rise and fall delays, which will very likely cause glitch hazards and unwanted power consumption.

In this paper, we propose a ratioed 3-2 compressor to resolve all the problems mentioned above. We also study what kind of scheme suits the implementation of an individual product term summation in the inner product computation. Although researchers have proposed a variety of compressors in the past to reduce the bits in the parallel multiplier array, such as 4-2, 5-5-4, or 9-2 compressors, etc. [11, 12], we still adopt the primitive 3-2, or 7-4 compressors instead of more advanced structures in our study for the following two reasons. First, most of the proposed compressors utilise the fact that the partial products are generated simultaneously, so that some of the input signals of the compressor are carried in from its right neighbour. The individual product term summation in the inner product computation can be treated as a special case of the partial-product reduction in the multiplication operation, in which only one single column of partial products is summed. Thus, no design of the compressors which assumes carry-in signals passed from its right neighbouring column is needed in the individual product term summation, since it might increase the height of the entire compressor tree. Secondly, Oklobdzija *et al.* [13]

IEE Proceedings online no. 20000382

DOI: 10.1049/ip-cdt:20000382

Paper received 30th September 1999

The authors are with the Department of Electrical Engineering, National Sun Yat-Sen University, Kaohsiung, Taiwan 80424

pointed out that it is the interconnection of the compressors rather than the structure of the compressors that leads to the fastest realisation of partial product reduction in the multiplication operation. The outperformance of the 4-2 compressors and 7-3 compressors built by Zhang *et al.* [10] verified that the conclusion of Oklobdzija *et al.* is correct. Therefore, in this paper, we will concentrate on the arrangement of the 3-2 compressors such that the carry propagation delay in the critical path is reduced. Three alternative architectures of the digital ratioed compressors built with the properly ordered 3-2 compressor building blocks are presented and compared. Analytical forms of carry propagation-delay estimation for these architectures are also respectively derived. The HSPICE and Verilog simulation results are also presented to verify the correctness of our observation. The final layout was compiled and delivered to the TSMC (Taiwan Semiconductor Manufacturing Company) to produce the physical chip. Finally, the real chip was fabricated and fully tested by an IMS digital tester of the ATS. The test results verify the correctness of our design.

2 Framework of ratioed compressors

2.1 Basic compressor building block design

A 3-2 compressor is basically a full adder. The feature of such a compressor is that the output represents the number of 1s given in inputs. The equations of a full adder are presented as follows:

$$\begin{aligned} S &= (a \oplus c)b' + (a \oplus c)'b = Fb' + F'b \\ C &= (a \oplus c)b + (a \oplus c)'c = Fb + F'c \end{aligned} \tag{1}$$

where F denotes $(a \oplus c)$.

As shown in Fig. 1, the logic structure of a typical 3-2 compressor can be split up into two logic layers. One of the three inputs, $b(b')$, is not required in the first logic layer. The existence of unequal delays in the 3-2 compressor paves the way for us to reduce the total delay of the inner product computation by arranging the input signals to the 3-2 compressors inside the compressor tree in a proper order.

2.2 Ratioed 3-2 compressor design

Although a 3-2 compressor can be realised by a full adder, and Zhang *et al.* [10] proposed a C^2PL design for 3-2 and 7-3 compressors, several design issues as addressed in Section 1 are still ignored in their work. Figs. 2 and 3 show the schematic diagrams for the two types of 3-2 compressors based on complex complementary pass-transistor logic (C^2PL) proposed in [10]. We used TSMC (Taiwan Semiconductor Manufacturing Company) 0.6 μm 1P3M technology to re-design the 3-2 compressors, and the schematic diagrams for the ratioed 3-2 compressors are shown in Figs. 3, 4 and 5. In Section 3 of this paper, we will demonstrate that the re-designed 3-2 compressor circuits overcome all the problems mentioned in Section 1.

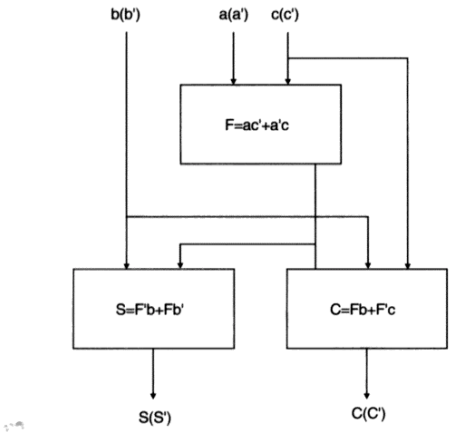

Fig. 1 *3-2 compressor building block*

2.3 The first architecture of ratioed compressors (architecture I)

A 7-3 compressor building block can be constructed by cascading four 3-2 compressors as shown in Fig. 6. A 15-4 compressor building block can also be formed similarly with two 7-3 compressors and two 3-2 compressors, as shown in Fig. 7. Based on this design methodology, an architecture for a $(2^n - 1)$-to-n compressor is composed of two $(2^{n-1} - 1)$-to-$(n - 1)$ compressors and $(n - 1)$ 3-2 compressors. Note that the logic layers required for the critical path can be reduced from 6 to 5 for the 7-3 compressor, shown as the dashed line in Fig. 6, if the inputs to the 3-2 compressor on the lower-left are properly ordered.

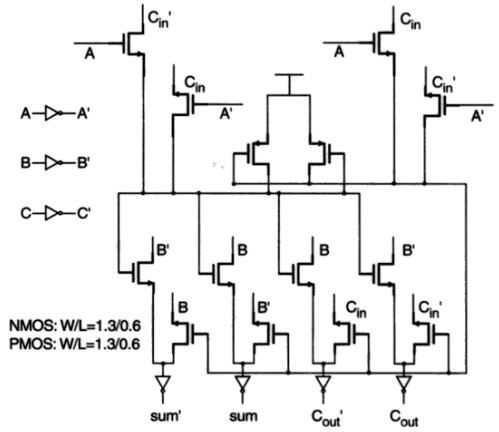

Fig. 2 *Schematic diagram for the C^2PL 3-2 compressor in Zhang's design [10], type 1*

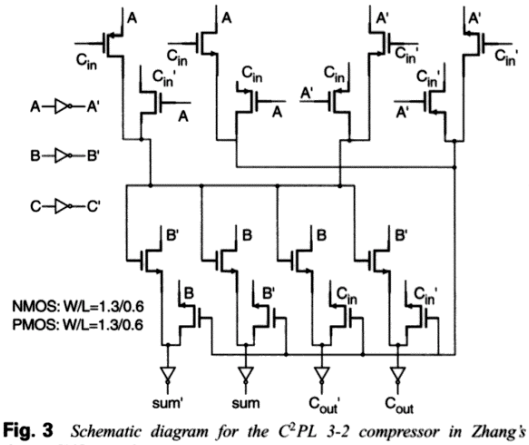

Fig. 3 *Schematic diagram for the C^2PL 3-2 compressor in Zhang's design [10], type 2*

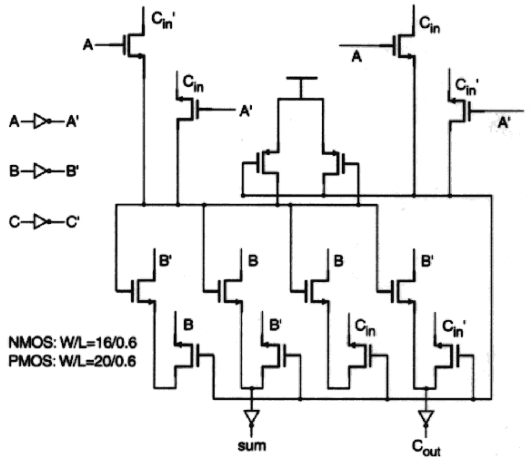

Fig. 4 *Schematic diagram for the re-designed C²PL 3-2 compressor, type 1*

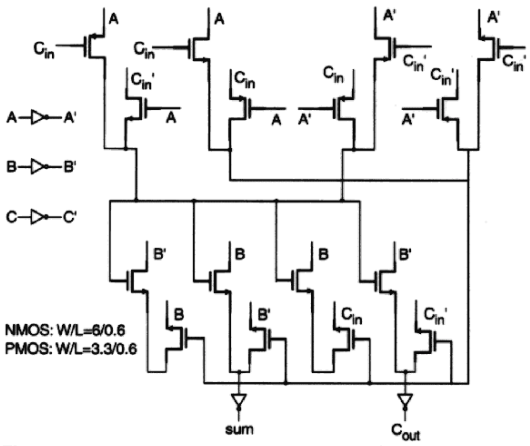

Fig. 5 *Schematic diagram for the re-designed C²PL 3-2 compressor, type 2*

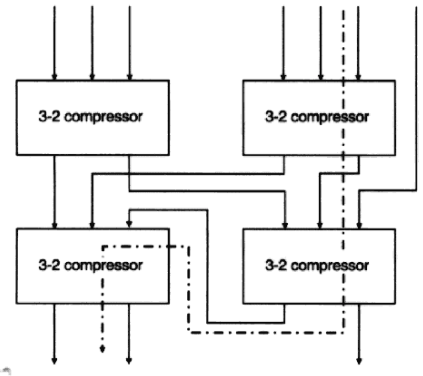

Fig. 6 *A 7-3 compressor building block*

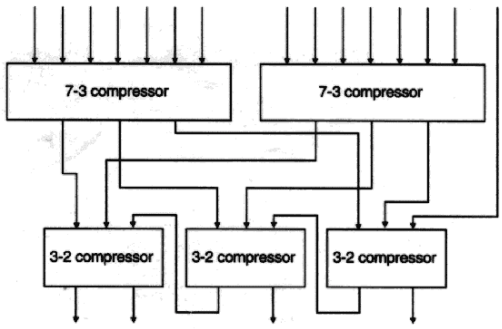

Fig. 7 *A 15-4 compressor for architecture 1*

2.3.1 Carry propagation delay equations: Since the total delay of such a design in approximately proportional to the count of the 3-2 compressors where the critical path resides, we assume $D1_n$ denotes the count of 3-2 compressors when $2^n - 1$ bits are fed into the $(2^n - 1)$-to-n compressor. By observing the structure of the compressor, we can deduce $D1_2$, $D1_3$ and $D1_n$ as follows:

$$D1_2 = 1$$

$$D1_3 = 3 = 1 + 2 = 2 + D1_2$$

$$D1_n = n - 1 + D1_{n-1} \qquad n \geq 3 \tag{2}$$

By solving the above recurrence relation, we obtain

$$D1_n = \frac{n(n-1)}{2} \tag{3}$$

Apart from the delay for the single building block, we have to count the processing stages needed for summing individual inner product term. The number of processing stages is roughly estimated as

$$\frac{\ln \frac{n}{M}}{\ln \frac{n}{2^n - 1}},$$

where n denotes the total bits of the basic building block output and M represents the bit count of the data inputs.

Therefore the count of the 3-2 compressors when M bits are fed into the $(2^n - 1)$-to-n compressor building blocks can be shown as follows:

$$D1_{M,n} \approx \frac{\ln \frac{n}{M}}{\ln \frac{n}{2^n - 1}} \cdot \frac{n(n-1)}{2} \tag{4}$$

Based on eqn. 4, a 3-D mesh of delay computation for n from 2 to 10 and M from 500 to 10000 can be built up as shown in Fig. 8.

As mentioned earlier, one of the three inputs goes through one logic layer, and the other two pass through both layers of the 3-2 compressor. By making proper connections globally among the 3-2 compressors inside a $(2^n - 1)$-to-n compressor, the total delay in the critical paths counted by the number of logic layers, $D2_n$ can be derived similarly as follows:

$$D2_2 = 2$$

$$D2_3 = 5 = 2 + 2 + 1 = D2_2 + 3$$

$$D2_n = D2_{n-1} + 3 \qquad n \geq 3 \tag{5}$$

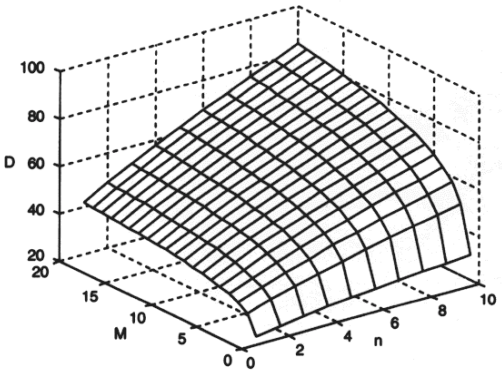

Fig. 8 *Delay of M data inputs using $(2^n - 1)$-to-n-compressors*

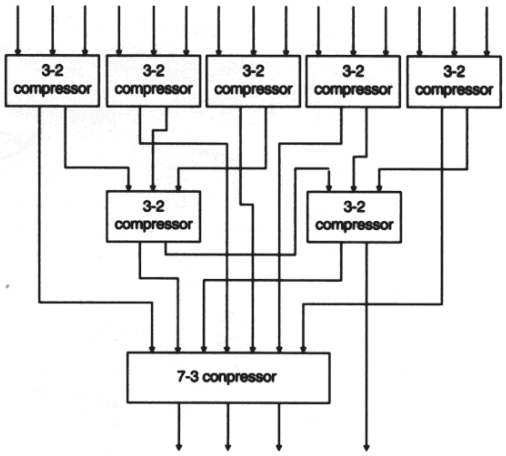

Fig. 9 *A 15-4 compressor for architecture II*

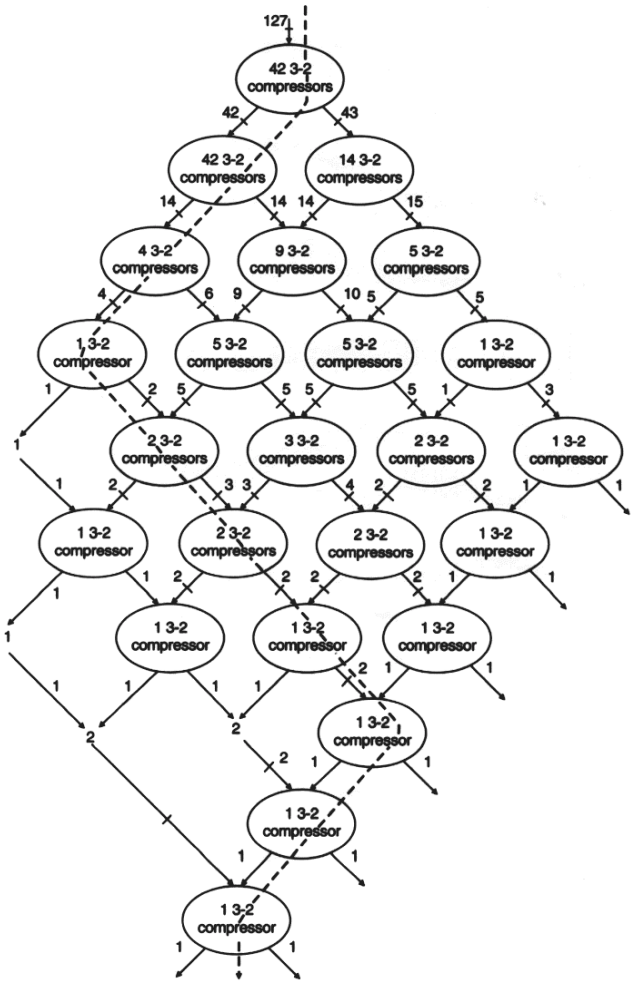

Fig. 10 *A 127-7 compressor for the systolic like architecture (architecture III)*

By solving the above recurrence relation, we obtain

$$D2_n = 3n - 4 \tag{6}$$

Note that $D2_n$ is counted by the number of logic layers, whereas $D1_n$ is counted by the block count of the 3-2 compressors, which is composed of two logic layers.

2.3.2 The count of 3-2 compressor building blocks: Based on the configuration of this architecture given in Fig. 7, we can derive the count of the 3-2 compressors used in a $(2^n - 1)$-to-n compressors, ξ_n, as follows:

$$\begin{aligned} &\xi_2 = 1 \\ &\xi_3 = 4 \\ &\xi_n = 2 \cdot \xi_{n-1} + n - 1 \qquad n > 2 \\ \Rightarrow\ &\xi_n = 2^n - n - 1 \end{aligned} \tag{7}$$

2.4 The second architecture of the ratioed compressors (architecture II)

Referring to Fig. 9, an alternative architecture of the 15-4 compressor building block can be constructed with five 3-2 compressors and one 7-3 compressor. In contrast to the first architecture, this architecture is composed of parallelised 3-2 compressor building blocks at the first processing stages and one single $(2^{n-1} - 1)$-to-$(n - 1)$ compressor at the last processing stage. This architecture apparently alleviates the lengthy carry propagation delay introduced by serialised connection of the 3-2 compressors shown in Fig. 5.

Although it is difficult to derive the analytical form of total delay of $(2^n - 1)$-to-n compressors for this architecture, the upper bound for the delay counted by the number of the 3-2 compressors, $D1_n$ still can be computed by observing the structure of this architecture.

The difference between $D1_n$ and $D1_{n-1}$ for $n > 3$ can be derived as follows:

$$\begin{aligned} D1_n - D1_{n-1} &\approx \left\lceil \frac{\log 2^n - 1}{\log 3} \right\rceil < \left\lceil n \cdot \frac{\log 2}{\log 3} \right\rceil \\ &= [0.631 \cdot n] < 0.8 \cdot n \\ \Rightarrow D1_n &< 0.8 \cdot n + D1_{n-1} \end{aligned} \tag{8}$$

Thus, the upper bound of the total delay counted by the number of 3-2 compressors becomes

$$\begin{aligned} D1_n &< 0.8 \cdot \left(\frac{n^2 + n}{2} - 6 \right) + D1_4 \\ \Rightarrow D1_n &< 0.4 \cdot (n^2 + n) - 1.2 \end{aligned} \tag{9}$$

where $D1_4$ is 6. Note that we only derive $D1_n$ for $n > 3$ because both architectures presented above are topologically equivalent when n is less than or equal to 3.

2.5 The third architecture of the ratioed compressors (architecture III)

The last architecture was propose to improve the carry propagation delay of the critical paths is shown in Fig. 10. This architecture, inspired by the design methodology of systolic arrays, consists of parallelised 3-2 compressor building blocks only at every processing stage. The total delay is also approximately proportional to the count of the 3-2 compressors where the critical path resides, shown as the dashed line in Fig. 10. The carry propagation delay of $(2^n - 1)$-to-n compressors, where $n = 2$ to 13, has been computed by a program and can be formulised as follows:

$$D1_n = \begin{cases} 2n - 3 & 2 \le n \le 6 \\ 2n - 4 & 7 \le n \le 13 \end{cases} \tag{10}$$

$$D2_n = \begin{cases} 3n - 4 & 2 \le n \le 6 \\ 3n - 3 & 7 \le n \le 10 \\ 3n - 2 & 11 \le n \le 13 \end{cases} \tag{11}$$

where $D1_n$ denotes the total delay of a $(2^n - 1)$-to-n compressor by the number of 3-2 compressor blocks, $D2_n$ indicates the total delay counted by the number of logic layers within a 3-2 compressor.

Based on eqn. 4, the upper bound for the delay counted by the number of the 3-2 compressors for this architecture, $D1_n$, can also be derived as follows:

$$\begin{aligned} D1_n &\approx \left\lceil \frac{\log \frac{2^n - 1}{2}}{\log \frac{3}{2}} \right\rceil < \left\lceil \frac{(n-1) \cdot \log 2}{\log \frac{3}{2}} \right\rceil \\ &< (n-1) \cdot \left\lceil \frac{\log 2}{\log \frac{3}{2}} \right\rceil < 2 \cdot (n-1). \end{aligned} \tag{11}$$

Compared with the other two architectures presented in the preceding subsections, the systolic like architecture improves the delay counted by the number of 3-2 compressors from $O(n^2)$ to $O(n)$. Apparently this outperformance is associated with the parallelised computing ability at each processing stage, as shown in Fig. 10.

In order to show the improved performance of this systolic like architecture in the carry propagation delay of the critical paths counted by the number of 3-2 compressors, Table 1 and Fig. 11 show the comparison with the other two architectures that we presented above.

3 Simulation analysis and chip implementation

3.1 Re-designed building blocks

In order to verify the correctness of our theoretical analysis, we used HSPICE and Verilog to conduct a series of simulations. The improvement of the asymmetrical rise

Table 1: Total delay comparison of the three architectures counted by the number of 3-2 compressors

Architecture \ n	2	3	4	5	6	7	8	9	10	11	12	13
I	1	3	6	10	15	21	28	36	45	55	66	78
II	1	3	6	10	14	19	25	31	38	45	52	60
III	1	3	5	7	9	10	12	14	16	18	20	22

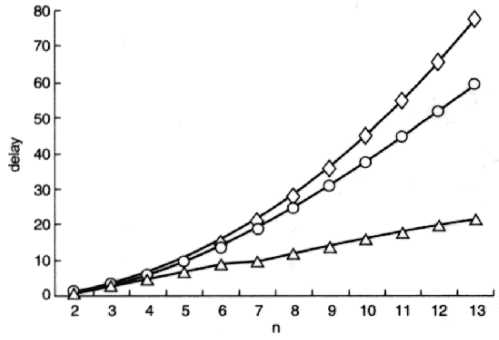

Fig. 11 *Total delay comparison of three architectures counted by the number of 3-2 compressors*

delay and fall delay in Zhang's design [10] can be illustrated through HSPICE simulations. The simulation results are tabulated in Table 2.

Figs. 12–15 show a comparison of the waveform diagrams of the original work in [10] and the re-designed 3-2 compressors. As shown in the Figures, the notoriously asymmetrical rise delay and fall delay appearing in Zhang's design [10] is fixed in our design. Besides some of the output of Zhang's 3-2 compressor cannot reach the full swing of voltage levels, which might be prone to noise problems.

3.2 Delay simulations

In Table 1, we showed theoretical comparisons among the carry propagation delays counted by the block levels,

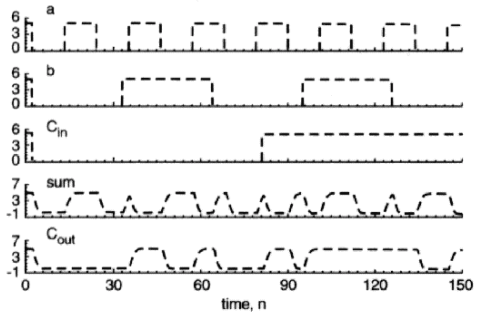

Fig. 12 *The waveform diagram of the C²PL 3-2 compressor in Zhang's design [10], type 1*

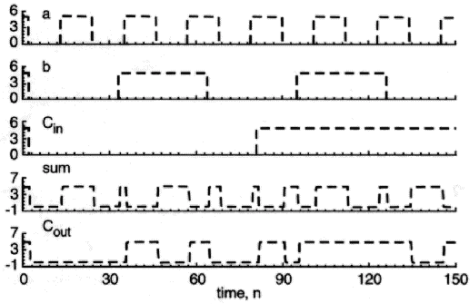

Fig. 13 *The waveform diagram of the re-designed C²PL 3-2 compressors, type 1*

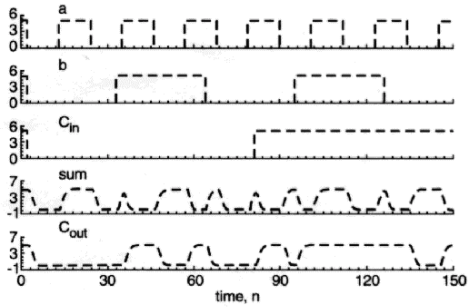

Fig. 14 *The waveform diagram of the C²PL 3-2 compressor in Zhang's design [10], type 2*

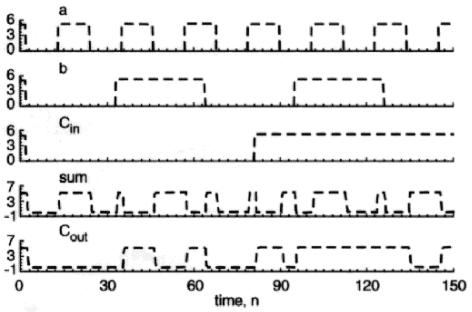

Fig. 15 *The waveform diagram of the re-designed C²PL 3-2 compressors, type 2*

Table 2: The comparison of the rise delay and the fall delay in Zhang's design and the re-designed 3-2 compressor

Circuit / Delay	Zhang's 3-2 compressor				Re-designed 3-2 compressor			
	C²PL(1)		C²PL(2)		C²PL(1)		C²PL(2)	
	carry	sum	carry	sum	carry	sum	carry	sum
Rise delay, ns	0.26	0.31	0.42	0.35	0.32	0.36	0.41	0.34
Fall delay, ns	0.87	0.83	0.87	0.87	0.24	0.43	0.39	0.42

Table 3: The comparison of the carry propagation delays for the three architectures

	15-4 compressor			31-5 compressor			63-6 compressor		
	architecture I	architecture II	architecture III	architecture I	architecture II	architecture III	architecture I	architecture II	architecture III
Delay, ns	2.8	2.8	2.8	4.1	4.2	3.9	5.6	5.5	4.8

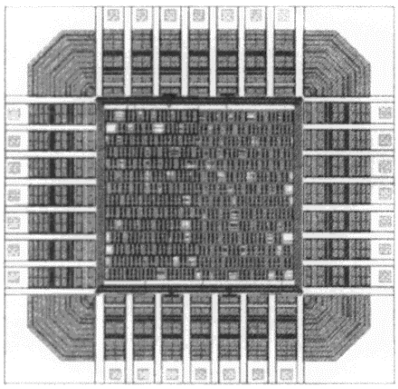

Fig. 16 *Circuit layout of the systolic like architecture of a 63-6 compressor*

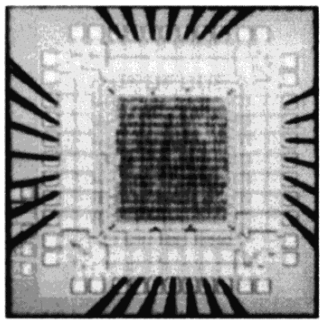

Fig. 17 *The die photo of the chip*

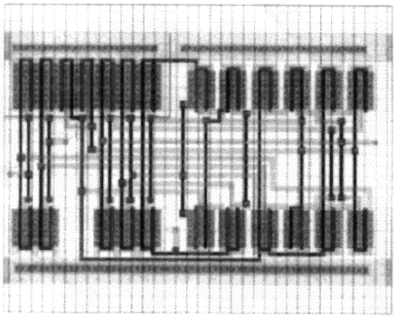

Fig. 18 *The circuit layout of a single cell for the C²PL 3-2 compressor, type 1*

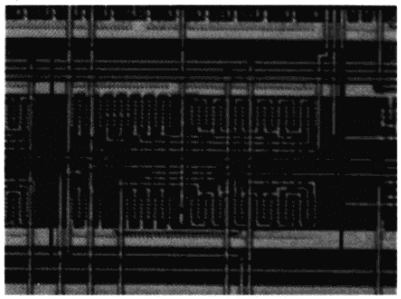

Fig. 19 *The die photo of a single cell for the C²PL 3-2 compressor, type 1*

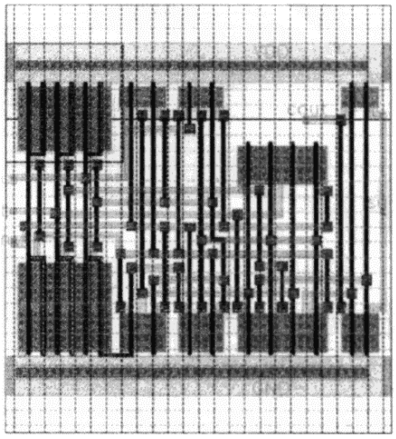

Fig. 20 *The circuit layout of a single cell for the C²PL 3-2 compressor, type 2*

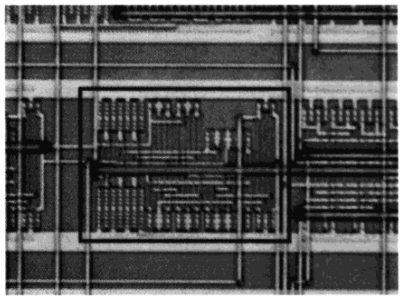

Fig. 21 *The die photo of a single cell for the C²PL 3-2 compressor, type 2*

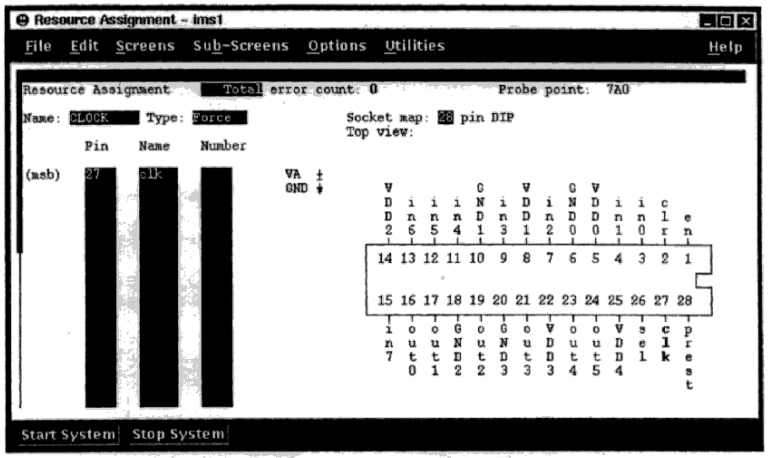

Fig. 22 *The physical pin-out setting*

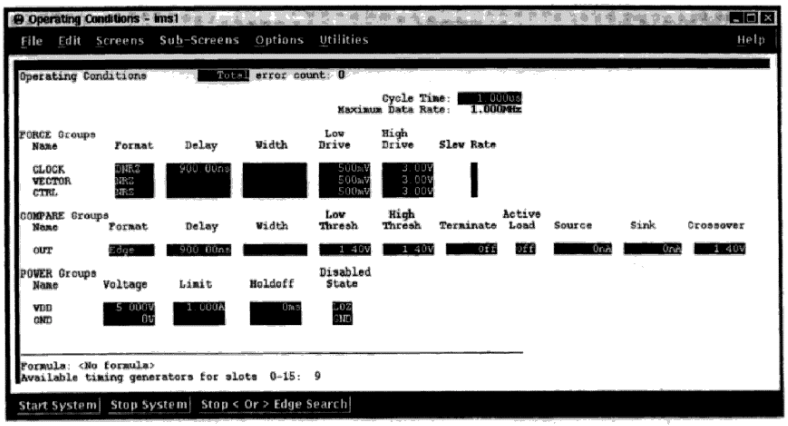

Fig. 23 *The I/O control signal setting*

which is the number of 3-2 compressors. of $(2^n - 1)$-to-n compressors for the three architectures that we proposed in Section 2. We then designed a series of compressors, which are 15-4, 31-5, and 63-6 compressors using the Verilog codes, and simulated them with 20000 random testing vectors, using the Verilog-XL simulator and its delay calculator, to explore the critical delays of the proposed three architectures. Table 3 illustrates the comparison of the carry propagation delays for the three architectures of the 15-4, 31-5 and 63-6 compressors when they are fed with 15, 31 and 63 data inputs, respectively.

The results reveals that the systolic like architecture of the digital ratioed compressors indeed produces the least carry propagation delay.

3.3 63-6 compressor chip implementation

The chip implementation of the systolic like architecture of the 63-6 compressor is done by using the Verilog HDL and COMPASS 1P3M cell libraries. Fig. 16 shows the circuit layout of the chip generated by Cadence Preview and Cell3 Ensemble. The real chip was tested by the IMS digital tester of ATS Co. The die photo of the chip is given in Fig. 17. Figs 18 and 19 show the circuit layout and die photo of a single cell for the $C^2PL(1)$ 3-2 compressor, respectively, while Figs. 20 and Fig. 21 present the circuit layout and die photo of a single cell for the $C^2PL(2)$ 3-2 compressor respectively. The physical pin-out and the test settings are illustrated in Fig. 22, while the I/O control signal settings are given in Fig. 23. A test comparison sample is given in Fig. 24 and a waveform sample is shown in Fig. 25. The maximum operating clock for the chip is up to 50 MHz.

4 Conclusions

In this paper, three compressors dedicated to the implementation of digital neural networks are presented and compared. The re-designed ratioed 3-2 compressor is proposed to overcome several problems appearing in Zhang's work in [10]. All the 3-2 compressors are arranged in a proper order to reduce the total delay in the critical path. The equations for counting the number of 3-2 compressors and logic layers in the critical path of a digital compressor are derived and used for exploration of the superior architecture of the digital ratioed compressor for

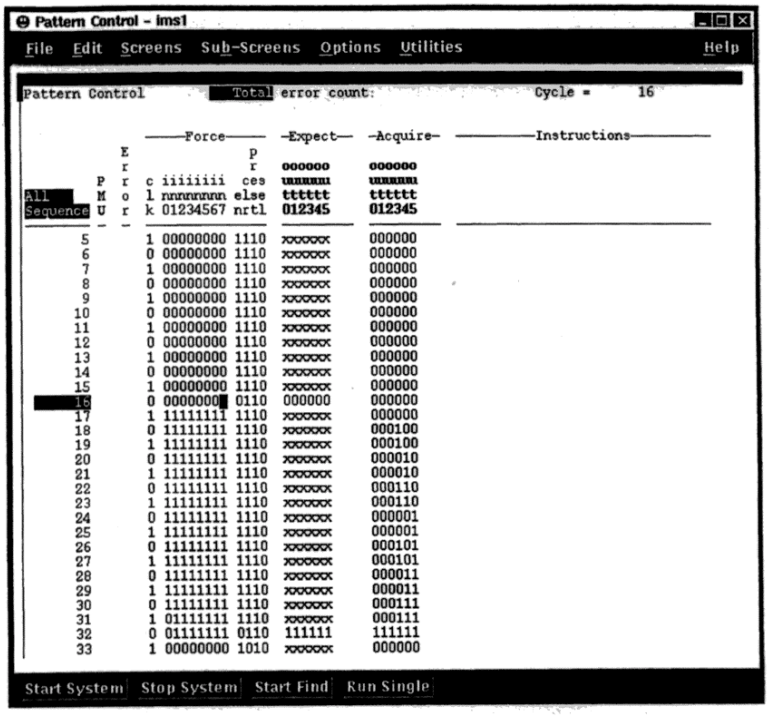

Fig. 24 *A test comparison sample*

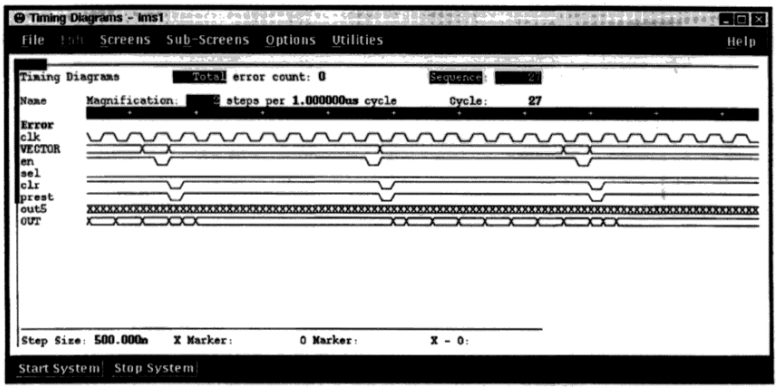

Fig. 25 *The waveform sample*

the digital neural-network applications. Our simulation results show that the systolic-like architecture gives a sub-optimal performance through the parallelised arrangement of 3-2 compressors at each stage of processing. Finally, we actually implemented the systolic-like architecture into a physical chip and proved it to be correct.

5 Acknowledgment

This research was partially supported by the National Science Council under grant NSC 88-2219-E-110-001

6 References

1 KOSKO, B.: 'Bidirectional associative memory', *IEEE Trans. Syst. Man Cybern.*, 1988, **18**, (1), pp. 49–60
2 WANG, C.-C., and DON, H.-S.: 'An analysis of high-capacity discrete exponential BAM', *IEEE Trans. Neural Netw.*, 1995, **6**, (2), pp. 492–496
3 KRAMBECK, R.H., LEE, C.M., and LAW, H.-S.: 'High-speed compact circuits with CMOS', *IEEE J. Solid-State Circuits*, 1982, **17**, pp. 614–619
4 GONCALVES, N.F., and DE MAN, H.J.: 'NORA: A race-free dynamic CMOS technology for pipelined logic structures', *IEEE J. Solid-State Circuits*, 1983, **18**, pp. 261–266
5 GU, R.X., and ELMASRY, M.I.: 'All-N-logic high-speed true-single-phase dynamic CMOS logic', *IEEE J. Solid-State Circuits*, 1996, **31**, (2), pp. 221–229
6 AFGHAHI, M.: 'A robust single phase clocking for low power high-speed VLSI application', *IEEE J. Solid-State Circuits*, 1996, **31**, (2), pp. 247–253

7 YUAN, J., and SVENSSON, C.: 'High-speed CMOS circuit technique', *IEEE J. Solid-State Circuits*, 1989, **24**, pp. 62–70
8 LEE, C.M., and SZETO, E.W.: 'Zipper CMOS', *IEEE Circuits Devices Mag.*, 1986, pp. 10–16
9 YANO, K., YAMANAKA, T., NISHIDA, T., SAITO, M., SHIMOHIGASHI, K., and SHIMIZU, A.: 'A 38-ns CMOS 16 × 16-b multiplier using complementary pass-transistor logic', *IEEE J. Solid-State Circuits*, 1990, **25**, (2), pp. 388–395
10 ZHANG, D., and ELMASRY, M.I.: 'VLSI compressor design with applications to digital neural networks', *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, 1997, **5**, (2), pp. 230–233
11 SONG, P.J., and DE MICHELI, G.: 'Circuit and archiecture trade-offs for high-speed multiplication', *IEEE J. Solid-State Circuit*, 1991, **26**, (9), pp. 1184–1198
12 STENZEL, W.J.: 'A compact high speed parallel multiplications scheme', *IEEE Trans. Comput.*, 1977, **262**, pp. 948–957
13 OKLOBDZIJA, V.G., VILLEGER, D., and LIU, S.S.: 'A method for speed optimized partial product reduction and generation of fast parallel multipliers using an algorithmic approach', *IEEE Trans. Comput.*, 1996, **45**, (3), pp. 294–305